FREQUENCY-WARPING INVARIANT FEATURES FOR AUTOMATIC SPEECH RECOGNITION

Alfred Mertins and Jan Rademacher

Signal Processing Group
University of Oldenburg, Department of Physics, 26111 Oldenburg, Germany
Email: {alfred.mertins, jan.rademacher}@uni-oldenburg.de

ABSTRACT

Based on the well-known relationship between vocal tract length (VTL) variation and linear frequency warping, we present a method for generating vocal tract length invariant (VTLI) features. These features are computed as translation invariant, correlation-type features in a log-frequency domain. In phoneme classification and recognition experiments on the TIMIT database, their discrimination capabilities and robustness to mismatches between training and test conditions turned out to be considerably better than for Mel-frequency cepstral coefficients (MFCCs). The best results are obtained when VTLI features and MFCCs are combined.

1. INTRODUCTION

The short-time spectra of two speakers A and B are approximately related as $X_A(\omega) = X_B(\alpha\omega)$, where α is the so-called warping factor. The value of α depends on the ratio of the vocal tract lengths of both speakers and usually lies in the range between 0.8 and 1.2, relative to an average speaker.

Vocal tract length normalization based on the above relationship has become an integral part of many automatic speech recognition engines [1, 2]. Recent approaches even normalize the utterances from the same speaker with optimal α on a frame-by-frame basis, in order to better match the standard realizations of the phonemes [3]. The value of α is typically selected as the one that yields the highest likelihood scores in a subsequent hidden Markov model (HMM) based recognizer [2, 3].

Besides warping of short-time spectra, also the computation of warping-invariant features has been proposed [4, 5]. In [4] the invariance was achieved with the scale transform, which has the property that the magnitude spectra of two signals $x(t)$ and $\frac{1}{\sqrt{\alpha}}x(t/\alpha)$ are the same. In [5], the continuous wavelet transform was used as a preprocessor, in order to obtain a speech representation in which linear frequency warping is converted to a translation in the log-frequency direction. In a second step, vocal tract length invariant (VTLI) features were generated by analyzing the wavelet representations in a translation-invariant manner. The methods studied in [5] include the auto- and cross-correlations of local wavelet spectra magnitudes as well as linear and nonlinear transforms thereof.

In this paper, we extend the work of [5] by looking at several ways of combining VTLI and classical MFCC features and analyzing the performance in various classification and recognition tasks. In particular, we look at the robustness of combined feature sets with respect to mismatches between the training and test conditions.

The paper is organized as follows. In the next section, we briefly discuss the wavelet transform from which warping-independent features are generated in a subsequent step. Section 3 then presents the proposed VTLI features. In Section 4 we describe the experimental setup and present results on phoneme classification and recognition experiments. Section 5 gives some conclusions.

2. THE WAVELET TRANSFORM

The wavelet transform of a continuous-time signal $x(t)$ is given by

$$\mathcal{W}_x(t,a) = |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} x(\tau)\, \psi^*\Big(\frac{\tau - t}{a}\Big)\, d\tau \qquad (1)$$

where $\psi(t)$ is the so-called mother wavelet, a is the scaling parameter, and the asterisk * denotes complex conjugation. By varying a, the center frequency, bandwidth, and effective time-width of $\psi(t/a)$ are changed according to the scaling theorem of the Fourier transform. This is known as a constant-Q analysis, where the relative bandwidth is constant. The wavelet transform $\mathcal{W}_{x_\alpha}(t,a)$ of a normalized, linearly frequency warped signal $x_\alpha(t) = \frac{1}{\sqrt{\alpha}}x(\frac{t}{\alpha})$, $\alpha > 0$, with Fourier spectrum $X_\alpha(\omega) = \sqrt{\alpha}X(\alpha\omega)$ is related to $\mathcal{W}_x(t,a)$ as $\mathcal{W}_{x_\alpha}(t,a) = \mathcal{W}_x\Big(\frac{t}{\alpha}, \frac{a}{\alpha}\Big)$.Thus, a linear frequency warping of the signal by a factor α results in a translation of the wavelet transform by $\log\alpha$ in the $(\log a)$-domain. This is important, because the wavelet transform is naturally computed for equally spaced values of $\log a$.

In order to compute the wavelet transform for a discrete-

This work was supported by the EU DIVINES Project under Grant No. IST-2002-002034.

time signal $x(n)$, we discretize (1) and use the definition

$$w_x(n,k) = 2^{-k/(2M)} \sum_m x(m)\, \psi^*\Big(\frac{m - nN}{2^{k/M}}\Big), \qquad (2)$$

where M is the number of voices (subbands) per octave, and N is the subsampling factor used to reduce the sampling rates in the wavelet subbands. Assuming K octaves, the scaling parameter a takes on values $a_k = 2^{k/M}$, $k = 0, 1, \ldots, MK-1$.

An important property of the wavelet transform (2) is that it is computed on a regular time grid with the same subsampling factor N applied to all frequency bands, regardless of the bandwidth. This is in contrast to the discrete wavelet transform (DWT), which operates on a dyadic grid and uses different sampling rates in different octaves. Due to the constant sampling rate in all frequency bands, the wavelet transform (2) does not suffer from the shift-invariance problem known from the DWT, provided that the factor N is sufficiently small and chosen in accordance with the number of voices (e.g., such that $N \leq M$). Rather than implementing (2) directly, which means a significant computational load, one may use the à trous algorithm [6], implemented separately for each of the M voices.

The wavelet analysis will have better time resolution at higher frequencies than needed for producing feature vectors every 5 to 15 ms. Direct downsampling of features will therefore introduce aliasing artifacts. Since we are mainly interested in the signal-energy distribution over time and frequency, we may take the magnitude of $w_x(n,k)$ and filter it with a lowpass filter in time direction before final downsampling. The final primary features will then be of the form

$$y_x(n,k) = \sum_\ell h(\ell)\, |w_x(nL - \ell, k)| \qquad (3)$$

where $h(\ell)$ is the impulse response of the lowpass filter, L is the downsampling factor introduced to achieve the final frame rate $f_s/(N \cdot L)$, and f_s is the sampling frequency. To avoid that the filtered values $y_x(n,k)$ can become negative, we assume a strictly positive sequence $h(n)$ like, for example, the Hanning window.

Fig. 1 gives an example of a wavelet analysis. In Fig. 1(b), which shows $|w_x(n,k)|$ as a grayscale image (i.e., a scalogram), the pitch is visible in most frequency bands. This pitch pattern is no longer visible in Fig. 1(c), which depicts $y_x(n,k)$.

3. WARPING-INVARIANT FEATURES

Due to the nature of $y_x(n,k)$, warping-invariant features can be easily generated by taking the Fourier transform of $y_x(n,k)$ with respect to parameter k and retaining only the magnitudes of the transform coefficients. However, this is only one of several possibilities to obtain warping-invariant features. Any feature extraction strategy that is independent of a translation with respect to k will allow us to achieve this goal.

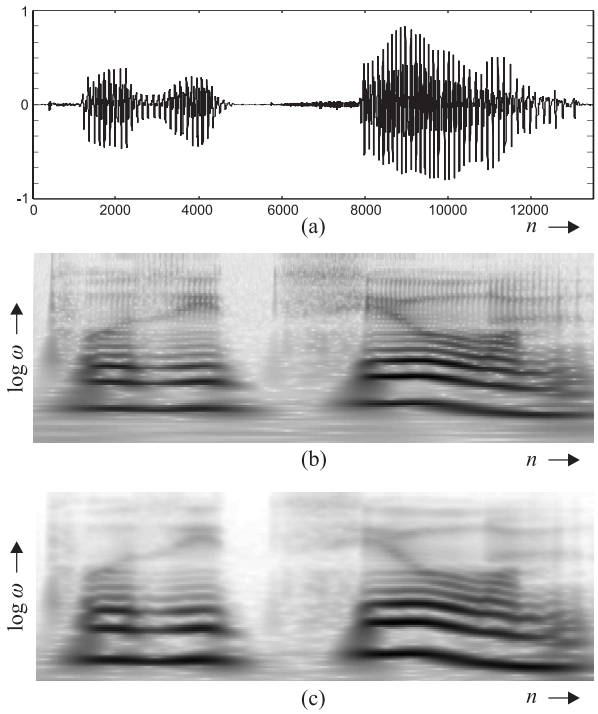

Fig. 1. Example of a wavelet analysis. (a) Time signal. (b) Wavelet transform magnitude $|w_x(n,k)|$ with $k \propto -\log\omega$. (c) Smoothed wavelet analysis $y_x(n,k)$.

Other possibilities include, but are not limited to correlation sequences between transform values or nonlinear functions thereof at two time instances n and $n-d$ (correlation with respect to the log-frequency index k). In particular, we here consider

$$r_x(n,d,m) = \sum_k y_x(n,k) y_x(n-d, k+m) \qquad (4)$$

and

$$c_x(n,d,m) = \sum_k \log(y_x(n,k)) \cdot \log(y_x(n-d,k+m)). \quad (5)$$

The parameter d is a time lag, and m is the lag for the log-frequency index k. The features $r_x(n,0,m)$ will give information on the signal spectrum in time frame n. For $d \neq 0$ the features $r_x(n,d,m)$ will give information on the development of short-time spectra over time. A feature vector for time index n can contain any collection of the above mentioned features computed for the same time index n. Moreover, any linear or nonlinear combination and/or transform or filtering of $r_x(n,d,m)$ and $c_x(n,d,m)$, including taking derivatives (i.e., delta and delta-delta features) will also yield warping invariant features.

To give an illustration of the properties of the correlation-based features, we consider the set $r_x(n,d,m)$ for $d = 0$ (i.e., autocorrelation features). Fig. 2 shows the features for the signal of Fig. 1. It is interesting to see that the autocorrelation, although it is in some sense phase-blind,

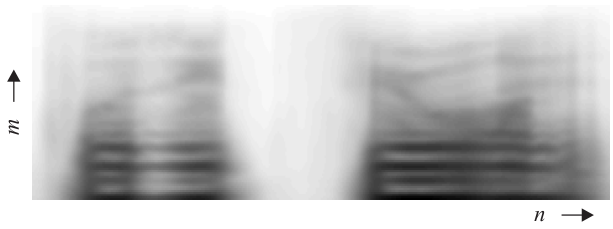

Fig. 2. Autocorrelation features $r_x(n, 0, m)$ for $m \geq 0$.

still retains the formant structure. This is due to the fact that noticeable correlation values are achieved when the high-energy pitch component is shifted and multiplied with the formant components during the correlation operation. Under the assumption that the linear warping model is true for vocal tract length variations, these formant-related structures will indeed be independent of the warping factor. For real speech, of course, this is only an approximation [7], but it leads to formant-like structures that are robust to vocal tract length variations.

4. EXPERIMENTAL RESULTS

In our experiments, a linear-phase wavelet transform based on the Morlet wavelet [8] given by $\psi(n) = \exp(j\omega_0 n) \times \exp(-\frac{n^2}{2\sigma_n^2})$ was used with the parameters $\omega_0 = 0.9\pi$ and $\sigma_n^2 = 100$. The transform $w_x(n, k)$ was carried out with $M = 12$ voices per octave and $K = 7$ octaves, resulting in primary feature vectors of length 84. The initial downsampling factor N was chosen as $N = 1$. The lowpass filter $h(n)$ was simply a rectangular window of 200 coefficients.

The original speech signals were sampled at 16 kHz sampling rate, and the final frame rate was set to 10 ms. The following 45 vocal-tract length invariant features (VTLI-F) were used:

- the first 20 coefficients of the discrete cosine transform (DCT) of $\log(r(n, 0, m))$ with respect to parameter m for $m = 0, 1, \ldots, 83$.
- the first 20 coefficients of the DCT of $c(n, 4, m)$ with respect to parameter m with $m = -83, \ldots, 83$.
- $\log(r(n, 4, m))$ for $m = -2, -1, \ldots, 2$

The warping-invariant features were also amended with classical MFCC features. For this, the 12 MFCCs and the single energy feature of the standard HTK setup were used (denoted by 13 MFCC in the following). Moreover, the first 15 DCT coefficients (DCT with respect to frequency parameter k) of the logarithmized wavelet features $\log(y_x(n, k))$ were used for feature set amendment as well. In addition, for all features, also the delta and delta-delta coefficients were included. Altogether, this makes a total number of 219 features. In a subsequent step, the number of features was reduced, using either feature selection, a linear discriminant analysis (LDA) [9], or combinations of the above.

The following feature sets were considered, where the factor 3 stands for the inclusion of delta and delta-delta features:

Table 1. Accuracies in % for phoneme classification.

Features	No. Feat.	Train.	Test	Acc.
3×13 MFCC	39	M+F	M+F	60.74
VTLI-F	39	M+F	M+F	61.33
VTLI-F+MFCC+WT	39	M+F	M+F	64.15
VTLI-F	47	M+F	M+F	62.97
3×13 MFCC + 8 VTLI-F	47	M+F	M+F	63.08
VTLI-F+MFCC+WT	47	M+F	M+F	64.80
3×13 MFCC + 3×5 VTLI-F	54	M+F	M+F	61.47
3×13 MFCC	39	M	F	50.01
VTLI-F	39	M	F	53.91
VTLI-F+MFCC+WT	39	M	F	57.00
VTLI-F	47	M	F	53.39
3×13 MFCC + 8 VTLI-F	47	M	F	53.92
VTLI-F+MFCC+WT	47	M	F	57.50
3×13 MFCC + 3×5 VTLI-F	54	M	F	52.54
3×13 MFCC	39	F	M	48.47
VTLI-F	39	F	M	53.75
VTLI-F+MFCC+WT	39	F	M	57.10
VTLI-F	47	F	M	52.96
3×13 MFCC + 8 VTLI-F	47	F	M	52.12
VTLI-F+MFCC+WT	47	F	M	57.25
3×13 MFCC + 3×5 VTLI-F	54	F	M	50.89

- 3×13 MFCC.
- VTLI-F: 3×45 VTLI features, reduced via an LDA to 39 and 47 features, respectively.
- VTLI-F+MFCC+WT: all 219 features, reduced via an LDA to 39 and 47 features, respectively.
- 3×13 MFCC + 8 VTLI-F: 3×13 MFCCs, amended with the 8 most important features from the plain VTLI-F setting above.
- 3×13 MFCC + 3×5 VTLI-F: MFCCs, amended with first five DCT coefficients of $\log(r(n, 0, m))$ with respect to the frequency lag m.

We present results for phoneme classification and recognition on the TIMIT corpus (including the SA files). The training and test sets were both split into male and female subsets in order to allow for training and testing under different conditions. In the following, M+F, M, and F denote training/test on male+female, male, and female data, respectively. Following the procedure in [10], 48 phonetic models were trained, and the classification/recognition results were folded to yield 39 final phoneme classes that had to be distinguished. The LDA was based on the 48 phonetic classes.

Table 1 shows results for maximum likelihood phoneme classification based on Gaussian mixture models with four mixtures and full covariance matrices. The models were built for single feature vectors, without further context. For each frame, the phonetic transcription that was valid for the frame center was used. For the M+F setting, where both male and female data was used during training and test, we see that the pure VTLI features (VTLI-F) yield almost the same performance as the MFCCs. However, when only male or only female data is used for training, the degradation for the VTLI features is far less than for the

Table 2. Correctness and accuracy in % for phoneme recognition using a HMM recognizer with five mixtures and diagonal covariance matrices. The definitions for correctness and accuracy are in accordance with the HTK documentation [11].

Features	No. Feat.	Training	Test	Correctness	Accuracy
3×13 MFCC	39	M+F	M+F	72.49	69.19
3×13 MFCC + 8 VTLI-F	47	M+F	M+F	73.68	70.74
VTLI-F+MFCC+WT	47	M+F	M+F	71.72	67.84
3×13 MFCC + 3×5 VTLI-F	54	M+F	M+F	72.90	69.33
3×13 MFCC	39	M	F	62.67	56.84
3×13 MFCC + 8 VTLI-F	47	M	F	67.68	62.27
VTLI-F+MFCC+WT	47	M	F	68.83	63.56
3×13 MFCC + 3×5 VTLI-F	54	M	F	65.76	59.38
3×13 MFCC	39	F	M	59.07	55.53
3×13 MFCC + 8 VTLI-F	47	F	M	63.33	60.79
VTLI-F+MFCC+WT	47	F	M	67.01	62.98
3×13 MFCC + 3×5 VTLI-F	54	F	M	63.13	59.13

MFCCs. The best performance is achieved when VTLI features, wavelet coefficients, and MFCCs are combined via an LDA to a final number of 47 features. This combined feature set is also the most robust one when the training and test conditions are different. Interestingly, the concatenation of the original MFCC features with some additional VTLI features improves the MFCC performance significantly. However, it cannot achieve the same robustness as the VTLI-F+MFCC+WT set. The results show the complementariness of invariant features and classical ones like MFCCs. Especially the robustness to a mismatch between training and test conditions is remarkable.

Table 2 presents results for HMM-based phoneme recognition using monophone models, three states per phoneme, five Gaussian mixtures per state, and diagonal covariance matrices. The recognizer was based on the Hidden-Markov-Toolkit (HTK). We see that the best performance is achieved when VTLI and classical spectral features are combined. The set 3×13+8VTLI-F gives the best performance when training and test conditions are the same, and it is also robust to a mismatch between training and test conditions. The highest robustness is achieved with the VTLI-F+MFCC+WT set, at the price of a small degradation when training and test conditions match.

5. CONCLUSIONS

We have proposed a technique for the extraction of vocal tract length invariant features. The performance of the new features has been demonstrated in both phoneme classification and recognition tasks. The results have shown that the new features are complementary to the well-known MFCCs and that they can be used to construct combined feature sets that are robust to speaker variations, especially when the training conditions do not match the test conditions. Future work will be directed toward finding more invariant features, investigating the noise robustness of the proposed features, and optimal feature combination.

6. REFERENCES

[1] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," in *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[2] L. Lee and R. C. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, Jan. 1998.

[3] A. Miguel, E. Lleida, R. Rose, L. Buera, and A. Ortega, "Augmented state space acoustic decoding for modeling local variability in speech," in *Proc. Interspeech 2005, Lisbon, Portugal*, 2005, pp. 3009–3012.

[4] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Scale transform in speech analysis," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 1, pp. 40–45, Jan. 1999.

[5] A. Mertins and J. Rademacher, "Vocal tract length invariant features for automatic speech recognition," in *Proc. 2005 IEEE Automatic Speech Recognition and Understanding Workshop*, San Juan, Puerto Rico, Nov. 27 -Dec. 1 2005, pp. 308–312.

[6] M. J. Shensa, "The discrete wavelet transform: Wedding the à trous and Mallat algorithms," *IEEE Trans. Signal Processing*, vol. 40, no. 10, pp. 2464–2482, Oct. 1992.

[7] G. Fant, "A non-uniform vowel normalization," *Speech Transmssion Lab. Rep., Royal Inst. Technol.,* Stockholm, Sweden, vol. 2-3, pp. 1–19, 1975.

[8] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1972.

[10] Kai-Fu Lee and Hsiao-Wuen Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 2, pp. 1641 – 1648, Nov. 1989.

[11] S. Young et al., *The HTK Book*, Cambridge University, 1995.